Quiz 1

1. According to the book Data Science for Business, ‘Data Science’ is only concerned with the automation of Machine Learning models.
Answer Key:**False**

2. Data should not be considered a strategic asset because it is difficult to attach monetary value to it.
Answer Key:**False**

3. Data Scientists don’t usually fall into a single bucket, in terms of the range of skills they bring to an organization.
Answer Key:**True**

4. Data Science problems are usually best solved by breaking the problem down into well-defined sub-problems.
Answer Key:**True**

5. Data cleaning is generally considered a low-value task so we should aim to avoid and minimize it.
Answer Key:**False**

6. Every data set can produce directly actionable insights if you analyze it thoroughly enough.
Answer Key:**False**

7. Classification and Regression are generally interchangeable terms in a data mining context.
Answer Key:**False**

8. We can perform unsupervised learning when we lack labels for the target variable in the training data.
Answer Key:**True**

9. Data “leakage” results in better training performance so is generally encouraged.
Answer Key:**False**

10. Select bias can hurt model generalizability so there are no circumstances in which we should allow selection bias to occur.
Answer Key:**False**

Quiz 2

1. Generally speaking, a “model” is an exact representation of reality, only smaller.
Answer Key:False

2. Induction is a philosophical concept that refers to generalizing from specific cases to general rules.
Answer Key:True

3. Entropy is minimized when all possible values of a discrete random variable have an equal probability of occurring.
Answer Key:False

4. Information gain measures the change in expected entropy due to partitioning a data set based on a specific variable's (or set of variables) values.
Answer Key:True

5. A Laplace correction is used to reduce the variance of probability estimates when sample sizes are small.
Answer Key:True

6. Decision Trees only produce decision boundary lines that are either parallel or perpendicular to an axis line.
Answer Key:True

7. Logistic regression is an example of a parametric model.
Answer Key:True

8. A Support Vector Machine is a linear separating hyper-plane that attempts to maximize the margin between positively and negatively labeled data points.
Answer Key:True

9. The squared-error loss function is less sensitive to large errors than the mean-absolute loss.
Answer Key:False

10. The magnitudes of the feature weights learned in a logistic regression procedure are invariant to the scale of the corresponding features.
Answer Key:False

Quiz 3

1. The best classification threshold (i.e., the number above which your classification is positive) for any classifier that outputs a probability P(Y|X) is 0.5
Answer Key:False

2. If false negatives are much more expensive than false positives, then one should favor a classifier’s recall over precision.
Answer Key:True

3. A classifier that always predicts the negative class can never achieve a 0-1 Loss of 99%
Answer Key:False

4. The prediction output of Logistic Regression, SVMs and Decision Trees can all be directly used for ranking but not necessarily for probability estimation.
Answer Key:True

5. The AUC metric is not base rate invariant
Answer Key:False

6. One can use a cost-confusion matrix and a confusion matrix to compute the expected value of a classifier’s prediction
Answer Key:True

7. The F1-score is the harmonic mean of Precision and Recall
Answer Key:True

8. The Precision of a classifier will never change if you down-sample the negative class in the training data.
Answer Key:False

9. Common training loss functions (MSE, Log-Loss, etc.) are good evaluation metrics when one cares about expected value estimation.
Answer Key:True

10. When two ROC curves cross one cannot conclude that a single model is better for all classification thresholds
Answer Key:True

Quiz 4

1. A Support Vector Machine is not designed to directly produce class probability estimates.
Answer Key:True

2. One should never use cross-validation if a dataset has more than a million records.
Answer Key:False

3. Model estimation bias is independent of sample size.
Answer Key:True

4. A “hyper-parameter” in a classification model is often a variable that controls the complexity of the underlying model.
Answer Key:True

5. When modeling with a proxy target variable, one should use a proxy that is correlated to the primary target variable, but has a lower base rate.
Answer Key:False

6. Given a single dataset, model estimation variance tends to increase as one uses algorithms with increased complexity.
Answer Key:True

7. If a data set has multiple instances of the same instance id (i.e., panel data), the best way to split into training and validation sets is to randomly sample the instances.
Answer Key:False

8. One must use the same loss function for model training as one uses for model validation.
Answer Key:False

9. When splitting data for model training and validation, one should always split by randomly partitioning the data set
Answer Key:False

10. Selecting an optimal subset of features from a large set is a form of model selection.
Answer Key:True